# 1.业务挑战

随着外资银行逐步进入竞争激烈的国内金融业，信用卡持卡客户的个性化要求越来越高等原因，使得中国信用卡市场的竞争日益白热化，而较高的客户流失率也逐渐成为各大银行不得不面对的一个现实问题。

在大量客户的基础上，如何根据现有的大数据，控制客户的流失并防止风险？

# 2.价值体现

通过客户信息提前预测到客户流失的可能性，可对用户做相应的客户回访，或其他产品推荐，已达到挽留客户降低用户流失率的目的。

# 3.解决方法概要

通过信用卡用户的相关信息分析，通过特征工程，选择出与客户流失直接或者间接的客户特征，把相关特征用于训练分类算法，然后用训练好的算法模型，预估现有用户的流失性。

# 3. Demo 链接，用户名和密码

demo地址IP：<https://9.30.109.52:9443>

在IBM公司内网的情况下，用chrom浏览器打开，用以下用户登录进入

username:tuser01

password:c6deshop

在Meetup工程下打开notebook: Churn\_analysis+training

**Note:firefox打开网页的话，可能会出错。如果不能看，也可以通过以下github:**

<https://github.ibm.com/ds-ml/bank-churn/blob/master/Churn_analysis%252Btraining.ipynb>

# 4. 数据文件介绍

1>Churn\_Rate.csv

有四列数据，分别为年份，第几季度，某年第几季度，用户流失率

数据内容截图如下：

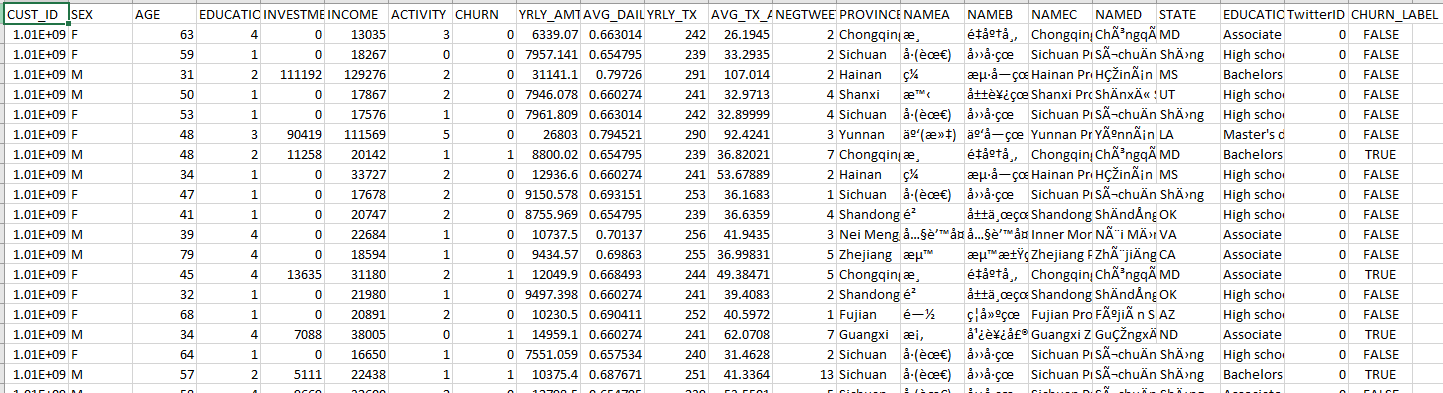
|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
| 2>CUST\_SUM.csv  从左往右数，每列内容依次为：  信用卡用户ID、性别（string）、年龄、教育程度（Int）、一年内投资的费用、收入、社区活动、是否流失（Int）、累计交易额、日均交易笔数、一年总交易笔数、平均每笔交易金额、在社交平台上对银行所做的评价量、所在地区(string)、教育程度(string）、Twitter账户、是否流失（bool）  注：教育程度分Int和string, 是同样的意思。是否流失分bool和Int类型，也是同样的意思。 |  |  |  |

3>CUST\_SUM\_CHINA.csv

从左往右数，每列内容依次为：

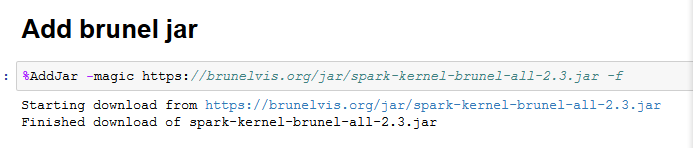
信用卡用户ID、性别（string）、年龄、教育程度（Int）、一年内投资的费用、收入、社区活动、是否流失（Int）、累计交易额、日均交易笔数、一年总交易笔数、平均每笔交易金额、在社交平台上对银行所做的评价量、省份（拼音），省份（简写汉字，如豫），省份（汉字全拼，如河南），省份（全拼拼音，同provice列），省份（带音调的拼音），所在地区(string)、教育程度(string）、Twitter账户、是否流失（bool）

注：教育程度分Int和string, 是同样的意思。是否流失分bool和Int类型，也是同样的意思。

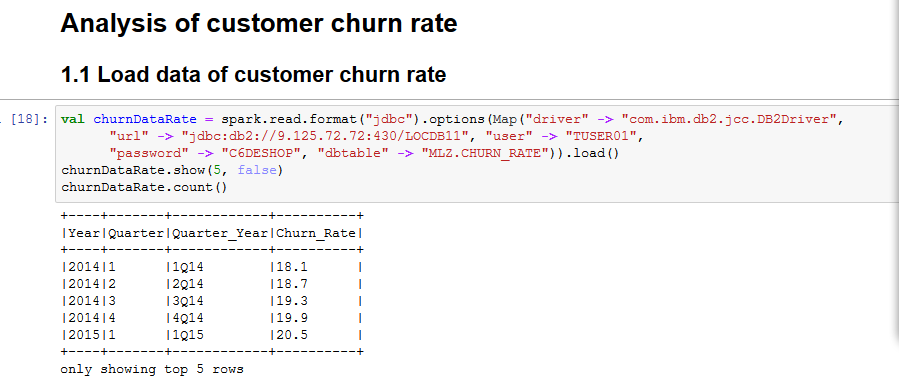


# 5. 详细步骤

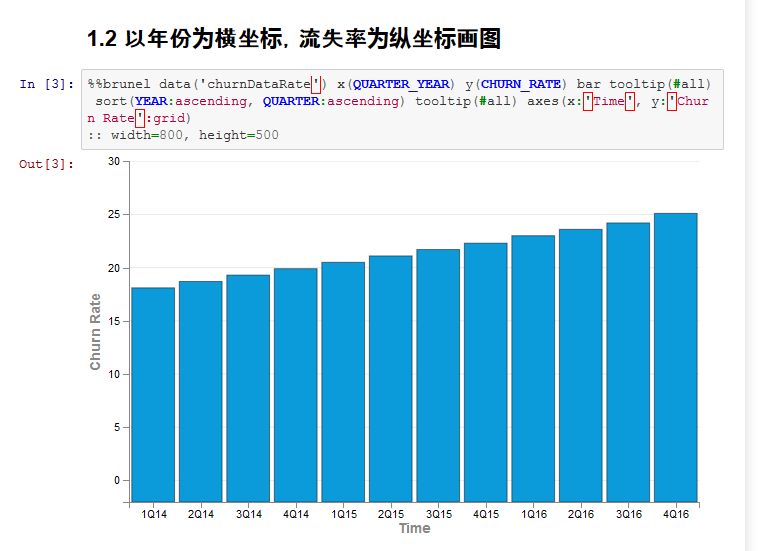
1>添加画图的包



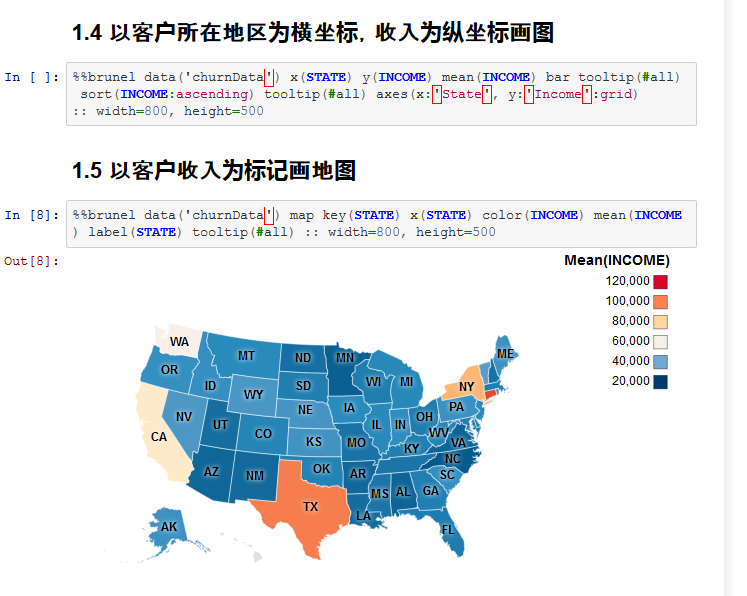
2>从DB2中加载数据，也可以把本地文件上传到ML for z上，再加载进来。



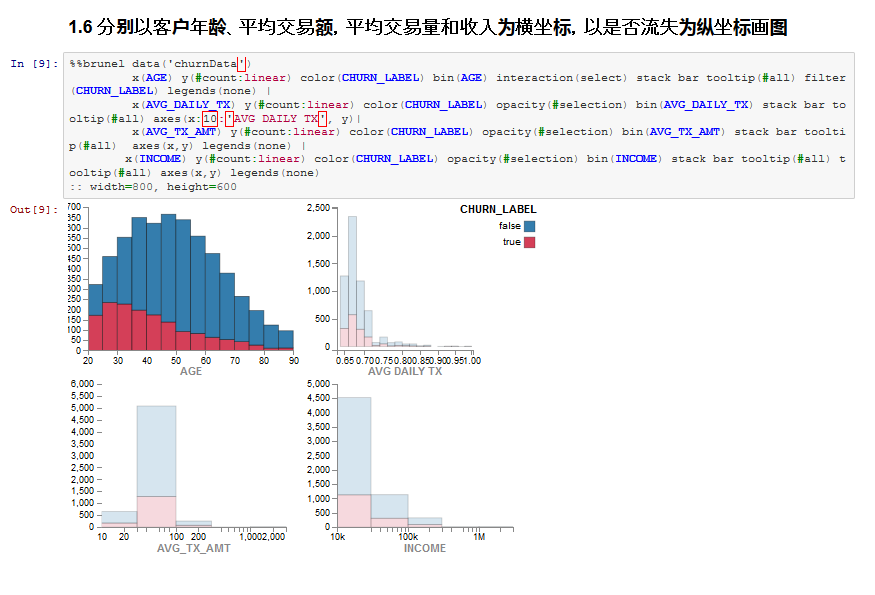
3>以年为横坐标，以流失率为纵坐标画图，可视化分析流失率的逐年变化。1Q14代表14年第一季度



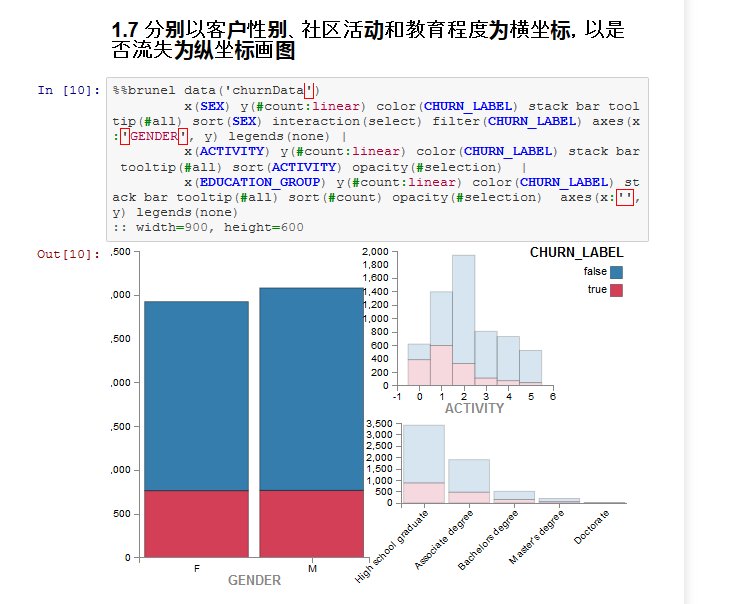
4.**以客户所在地区为横坐标，收入为纵坐标画图**。分析客户收入与地区的关系。1.5的是从地图的角度来看客户收入与地区的关系，从label来看，颜色越红的地区，收入越高。从地图中明显可以看出TX>NY>CA>WA>其他美国的州



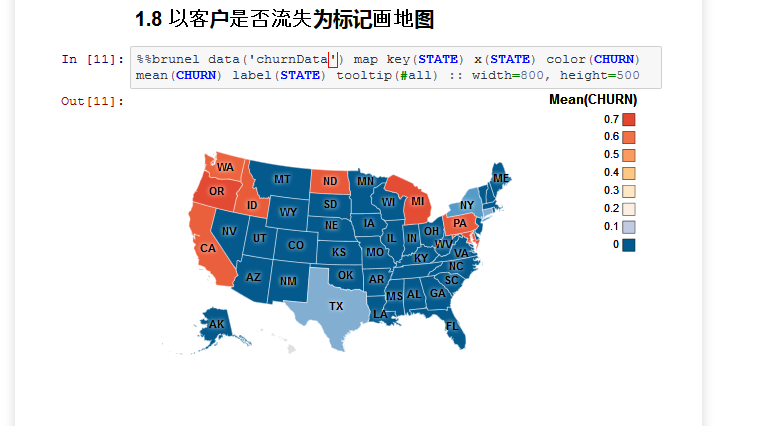
5.同理，画柱状图分析**客户年龄、平均交易额，平均交易量和收入与客户是**否流失的关系，从图中可以看出随着横坐标的变化，流失率有变化，说明这些特征与客户流失是相关的。



6.同理，画柱状图分析**客户性别、社区活动和教育程度是否流失的**关系，从图中可以看出随着横坐标的变化，流失率有变化，说明这些特征与客户流失是相关的。



7.从地图的角度分析客户流失是否跟地区有关系，由图可知，并不是收入最高的州流失率越大，跟客户所在地区有一定关系。

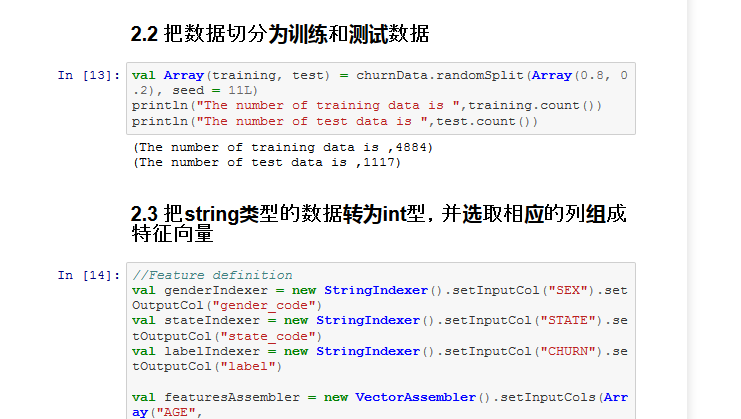


8>导入模型训练相应的库，比如说逻辑回归算法，pipeline，交叉验证，以及参数设置的包和二分类算法评估包

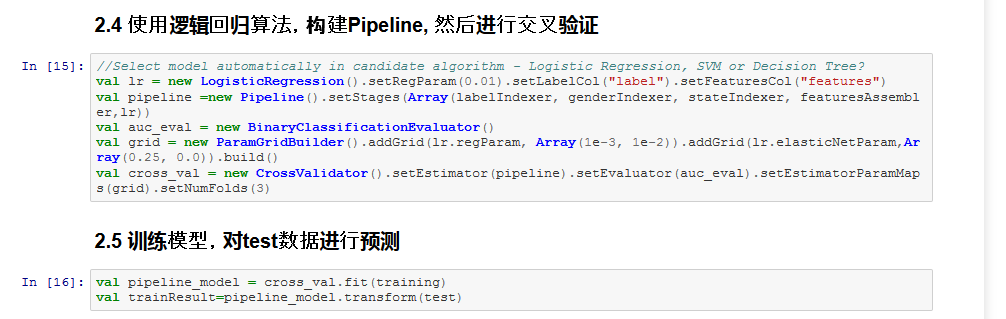


9>将数据切分成两部分，测试集与训练集占比0.8,0.2. 为了便于用算法进行训练，将相应string类型的数据变成从1-n的数字类型。

并选取年龄、社区活动、教育层次、在社交网站的负面评论、收入、性别、所在地区作为特征，用于以下训练算法。’



10>选取逻辑回归算法，并设置参数。构建pipeline、交叉验证模型。pipeline便于调参，将把string类型的特征变整形、选取特征、构建LR算法进行流水线操作。将lr的regparam参数固定在（1e-3,1e-2）之间，将lr的elasticNetParam参数固定在（0.25，0）之间变化。然后取三折十字交叉验证防止过拟合。



11>**评价模型，并画出ROC curve**

